Performance Evaluation of Country-Data using Unsupervised and Supervised Models Optimized by Genetic Algorithms

Nathan Ghenassia[1], Ónar Alberto García Hernández[2], Darryel Brown Murillo[3], and Juan Murillo-Morera[4]

[1,2,3,4]Lead University Costa Rica
Email: {nathan.ghenassia, onar.garcia, darryel.brown, juan.murillo.morera}@ulead.ac.cr

Abstract—Data science offers a powerful toolkit for understanding global socioeconomic disparities. This study investigates the effectiveness of a combined modeling approach for identifying countries in high need of assistance, utilizing a country-level dataset sourced from Kaggle. The proposed framework leverages unsupervised clustering, supervised classification and regression analysis. By integrating insights from all models, the framework identifies countries exhibiting characteristics of low development, high aid requirements, and belonging to clusters with similar disadvantaged profiles. This data-driven approach, informed by a rich Kaggle dataset, advances the field of data science for global development. It demonstrates the value of combining best-in-class models and genetic algorithms to leverage the quality and efficiency of this analysis process, ultimately informing policy interventions aimed at promoting global equity and sustainable development.

Index Terms—Unsupervised, Supervised, Classification, Regression, Learning Algorithms, Cluster, Prediction, Classification, Genetic Algorithm.

I INTRODUCTION

In the modern era, the explosion of data and the omnipresence of technology have radically transformed the way we understand and address social and economic issues on a global scale. In this context, data science has emerged as a powerful tool that enables the analysis, interpretation, and extraction of meaningful knowledge from complex and heterogeneous datasets [1][2]. One field where the application of data science has shown significant potential impact is in understanding global socioeconomic dynamics.

Access to economic and social data from multiple countries has allowed researchers and experts in the field of data mining to explore and better understand the challenges and disparities faced by different nations [3][2]. In particular, the ability to perform clustering through unsupervised models has become an invaluable tool for identifying underlying patterns and trends in data, thereby enabling the identification of countries with specific and prioritized needs [1][2].

In this context, the present work focuses on the application of data mining techniques to address the crucial question of identifying and understanding socioeconomic disparities among countries worldwide. The main objective is to use clustering methods to classify countries into different categories according to their economic, social, and educational situation. Through this approach, the aim is to identify those countries facing greater challenges and needs, as well as those showing outstanding performance in various socioeconomic metrics.

The problem addressed by this study is of vital importance in the field of international politics and global development. Understanding disparities among countries and identifying those requiring greater attention and support can help effectively guide international aid efforts, resource allocation, and policy implementation aimed at promoting sustainable development and reducing inequalities.

As we navigate the intricate web of global socioeconomic dynamics, the utilization of data science methodologies stands as a beacon of hope in our quest for understanding and addressing disparities among nations [2][3]. By harnessing the power of data analysis and advanced clustering techniques, this study endeavors to shed light on the multifaceted challenges and opportunities that define our world. Through a nuanced examination of economic, social, and educational indicators, we strive not only to identify those countries in greatest need but also to pave the way for evidence-based policy interventions aimed at fostering inclusive development and reducing inequality on a global scale [3]. In doing so, we reaffirm the transformative potential of data-driven insights and the pivotal role of data science in shaping a more equitable and prosperous future for all [2][3].

Conforming to the Goal-Question-Metric (GQM) paradigm [2], the goal of this research can be stated as follows:

Apply: Unsupervised, Supervised Classification and Regression models

For the purpose of: cluster, classificate and predict variables

With respect to: one dataset (Country-data)

From the point of view of: researchers and data science practitioners

In the context of: finding countries that are in the direst need of assistance.

The structure of the rest of the article is as follows. Section II presents research questions. Section III discusses related and state-of-the-art works. The proposed framework is explained in Section IV. Section V contains mathematical modeling. Section VI reports the performance of learning algorithms and hypotheses, and statistical tests. Finally, Section VII discusses conclusions.

II RESEARCH QUESTIONS

This section lists the main research questions that we set out to answer. Our analysis considered one dataset (Country-data):

- RQ-1 Which insights can be derived from the dataset using Unsupervised clustering models?
- RQ-2 Which Supervised classification models provide the best performance for predicting outcomes in the dataset?
- RQ-3 Which Supervised regression classification models provide the best performance for predicting outcomes in the dataset?
- RQ-4 Which countries are in the direst need of assistance?

III RELATED WORK

The following works have used Unsupervised, Supervised Classification and Regression learning algorithms for their studies:

In 2020, Alloghani et al. [5] conduct a Systematic Review on Supervised and Unsupervised Machine Learning Algorithms: Although this paper covers both supervised and unsupervised learning, it provides valuable insights into the latter. It discusses various unsupervised machine learning algorithms, including clustering, association, and dimensionality reduction.

In 2021, Baboshkin et al. [7] challenged traditional country classifications using a ”non-classical” approach with machine learning clustering algorithms. They used the silhouette score to evaluate cluster quality. Their results showed an optimal number of 5 clusters with a silhouette score of 0.68, indicating a good cluster structure. The study identified distinct groups of countries based on their open innovation indicators, providing a more nuanced classification compared to traditional economic indicators.

In 2023, Jerin et al. [8] conducted a comparative study of clustering algorithms for crime analysis. They evaluated six clustering algorithms (K-Means, DBSCAN, Hierarchical, Gaussian Mixture, Mean Shift, and OPTICS) using silhouette score and Calinski-Harabasz index. K-Means performed best with a silhouette score of 0.71 and Calinski-Harabasz index of 1243.5, followed closely by Hierarchical clustering.

In 2023, Sohan et al. [6] proposed a data-driven approach using K-means clustering and multidimensional poverty indices. They used the Hopkins statistic (0.76) to confirm the clustering tendency of their data. The silhouette score for their optimal clustering solution was 0.68, indicating good cluster separation. Their approach identified four distinct clusters of countries with varying levels of development needs.

In 2024, Sabouri et al. [4] conducted a comparative Study of Supervised Regression Algorithms in Machine Learning: This study delves into the performance evaluation of various supervised regression algorithms. While the paper primarily focuses on regression, it provides insights into the comparative performance of different supervised learning techniques.

In 2016, Murillo-Morera et al. [18] conducted an empirical evaluation of NASA-MDP datasets using a genetic defect-proneness prediction framework. They analyzed 864 learning schemes combining various data preprocessing techniques, attribute selectors, and learning algorithms. Their genetic algorithm-based approach demonstrated stability across different versions of datasets with varying noise levels. The framework achieved AUC values ranging from 0.7124 to 0.8383 in the evaluation phase and 0.7145 to 0.8554 in the prediction phase. For the best performing combinations, they found that BoxCox transformation (α=0.05) for data preprocessing, Linear Forward Selection or Forward Selection for attribute selection, and Bagging or LogitBoost for learning algorithms were most effective. This study showcases the potential of genetic algorithms in optimizing learning schemes for defect prediction, which is highly relevant to our study's use of genetic algorithms for model optimization in country data analysis.

These studies lay the groundwork for our research, which aims to use data (clustering, classifying, predicting) to understand which countries need the most help. It builds on past research and uses a big dataset about countries. This will help us figure out which methods work best to identify countries needing aid, ultimately leading to better ways to distribute resources around the world.

IV PROPOSED FRAMEWORK

A. Data Set

To carry out these experiments, we utilize the "Country-data.csv" dataset, which contains relevant information about various countries. The dataset consists of 167 countries (observations) and 10 variables, namely: country, childmort, exports, health, imports, income, inflation, lifeexpec, totalfer, gdpp.

B. Learning Scheme

As indicated in Section I of our study, three main learning approaches were utilized: Unsupervised, Supervised Classification, and Regression, each with its respective methods and techniques. The learning schemes consist of four parts: Data Preprocessing, Attribute Selector, Machine Learning Algorithm (referred to as the learning algorithm) and Genetic Algorithm.

Data Preprocessing (DP): The initial step in our learning scheme involved data preprocessing to ensure the quality and consistency of the input data. This phase included several tasks such as handling missing values, normalizing numerical features, encoding categorical variables, and scaling the data to standardize the range of independent variables. We employed techniques like mean imputation for missing values and Min-Max scaling to bring the data within a specific range, which is crucial for models like KNeighbors and algorithms that rely on distance measures.

Attribute Selector (AS): To improve the efficiency and performance of the models, we implemented an attribute selection process. This involved identifying and retaining the most relevant features from the dataset, thereby reducing dimensionality and eliminating irrelevant or redundant information. We used feature importance measures derived from tree-based models like RandomForest, and we also applied Recursive Feature Elimination (RFE) to systematically remove less significant features. The selection process was critical for enhancing model accuracy and reducing computational complexity, particularly when combined with the optimization capabilities of the genetic algorithm.

Learning Algorithms (LA): LinearRegression, DecisionTreeRegressor, RandomForestRegressor, Lasso, Ridge, KNeighborsRegressor, XGBRegressor. KNeighborsClassifier, DecisionTreeClassifier, RandomForestClassifier, AdaBoostClassifier, XGBClassifier.

Genetic Algorithms (GA): The Genetic Algorithm played a pivotal role in optimizing the performance of both classification and regression models. GA is a metaheuristic inspired by the process of natural selection, where it iteratively evolves a population of candidate solutions toward an optimal or near-optimal solution. For our study, the GA was used to optimize hyperparameters of the learning algorithms, such as the number of estimators in RandomForest, the maximum depth of DecisionTree, and the regularization parameters in Lasso and Ridge regression. By simulating evolutionary processes such as selection, crossover, and mutation, the GA was able to navigate the hyperparameter space effectively, leading to models that achieved higher accuracy and lower error rates compared to those optimized by traditional grid or random search methods.

C. Experimental Design

The objective of this study is to conduct a comprehensive analysis of a dataset using techniques of exploratory analysis, dimensionality reduction, unsupervised learning, classification and regression.

The experimental design was adopted following steps cited in the literature. The following steps were performed [18]:

1) We will use a dataset named "Country Data," which consists of 168 rows and 10 columns.

a) The dataset will be loaded using the 'pd.readcsv()' method from the Pandas library to read a CSV file.

b) Exploratory data analysis (EDA) will be conducted using methods such as 'describe () ' to obtain descriptive statistics of the Data Frame. Data distributions will be visualized using boxplots ('boxplot () '), density plots ('plot(kind='density') '), and histograms ('plot(kind='hist') '). Important to add that correlation between variables will be calculated using the 'corr () ' method. The correlation matrix will be visualized as a heatmap using the 'heatmap () ' method.

c) Regarding dimensionality Reduction, Principal Component Analysis (PCA) will be implemented using the 'PCAPrince () ' method to reduce the dimensionality of the data.

d) Within the Unsupervised Learning agglomerative hierarchical clustering methods such as 'ward () ', 'average () ', 'single () ', and 'complete () ' will be applied to cluster the data. Clustering methods such as 'KMeans () ' and 'KMedoids () ' will be used for clustering with K-Means and K-Medoids, respectively. Visualization methods such as 'TSNE () ' and 'UMAP () ' will be employed to visualize high-dimensional data in a two-dimensional space.

e) We used several classification models such as KNeighborsClassifier, DecisionTreeClassifier, RandomForestClassifier, AdaBoostClassifier, XGBClassifier, will be trained. The performance of each model will be evaluated using metrics such as overall accuracy.

f) Regression models will be implemented to predict continuous values based on independent variables. Linear Regression, which fits a linear regression model using multiple independent variables to predict the dependent variable. Lasso, this utilizes Lasso regression to fit a model and perform feature selection, applying L1 regularization to penalize coefficients of less important features. Finally, Ridge fits a Ridge regression model to control model complexity, applying L2 regularization to penalize coefficients of larger features.

g) For the experiment evaluation, metrics such as R-squared will be calculated to evaluate the accuracy of the regression models.

V MATHEMATICAL MODELING

V.1 Main Components

V.1.1 Generator

The generator G processes and prepares data iteratively:

$$G : \mathcal{DB} \xrightarrow{T_1} \mathcal{DB}' \xrightarrow{T_2} \mathcal{DB}'' \xrightarrow{T_3} (\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}) \tag{1}$$

where:

- $\mathcal{DB}$ is the original database
- $T_1 : \mathcal{DB} \to \mathcal{DB}'$ is the feature selection transformation
- $T_2 : \mathcal{DB}' \to \mathcal{DB}''$ is the target variable selection transformation
- $T_3 : \mathcal{DB}'' \to (\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}})$ is the train-test split, including standardization

V.1.2 Evaluator

The evaluator E incorporates the genetic algorithm to optimize hyperparameters iteratively:

$$E : \Theta \times \mathcal{D}_{\text{train}} \xrightarrow{\text{GA}_1} (\theta_1, \mathcal{M}_1) \xrightarrow{\text{GA}_2} (\theta_2, \mathcal{M}_2) \xrightarrow{\text{GA}_3} \cdots \xrightarrow{\text{GA}_G} (\theta^*, \mathcal{M}^*) \tag{2}$$

where:

- Θ is the hyperparameter space
- GA_i represents the i-th generation of the genetic algorithm
- θ_i are the optimized hyperparameters in generation i
- $\mathcal{M}_i$ is the optimized model in generation i
- $(\theta^*, \mathcal{M}^*)$ are the optimal hyperparameters and model after G generations

V.1.3 Predictor

The predictor P applies the optimized model to make predictions and evaluates performance:

$$P : \mathcal{M}^* \times \mathcal{D}_{\text{test}} \to (\hat{\mathcal{Y}}, \text{Performance}) \tag{3}$$

where:

- $\hat{\mathcal{Y}}$ is the set of predictions
- Performance is the evaluation metric (R-squared for regression, accuracy for classification)

V.2 Genetic Algorithm for Hyperparameter Optimization

For each model $m \in \mathcal{M}$, where $\mathcal{M}$ is the set of all models (regression or classification), the genetic algorithm optimizes the hyperparameters as follows:

$$\theta_m^* = \text{GASearchCV}(m, \Theta_m, \mathcal{D}_{\text{train}}) \tag{4}$$

The GASearchCV process :

$$\begin{aligned}
\text{GASearchCV :} &\text{Initialize population } P_0 = \{\theta_1, \theta_2, \ldots, \theta_n\} \text{ where } \theta_i \in \Theta_m && (5)\\
&\text{For generation } t = 1 \text{ to } G : && (6)\\
&\quad \text{Evaluate fitness: } f(\theta) = \text{CrossValidation}(m_\theta, \mathcal{D}_{\text{train}}) \text{ for } \theta \in P_{t-1} && (7)\\
&\quad \text{Select parents: } S_t = \text{TournamentSelection}(P_{t-1}, k) && (8)\\
&\quad \text{Create offspring: } O_t = \text{Crossover}(S_t) \text{ with probability } p_c && (9)\\
&\quad \text{Mutate offspring: } M_t = \text{Mutation}(O_t) \text{ with probability } p_m && (10)\\
&\quad \text{Update population: } P_t = \text{Elitism}(P_{t-1}) \cup M_t && (11)\\
&\text{Return } \theta^* = \arg\max_{\theta \in P_G} f(\theta) && (12)
\end{aligned}$$

where:

- $G = 8$ is the number of generations
- $n = 10$ is the population size
- $k = 5$ is the tournament size
- $p_c = 0.8$ is the crossover probability
- $p_m = 0.1$ is the mutation probability
- CrossValidation performs 5-fold cross-validation
- Elitism preserves the best individual from the previous generation

The fitness function $f(\theta)$ is defined as:

$$f(\theta) = \begin{cases} \text{R-squared score} & \text{for regression models} \\ \text{Accuracy} & \text{for classification models} \end{cases} \tag{13}$$

V.3 Model Evaluation

V.3.1 For Regression (R-squared)

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{14}$$

V.3.2 For Classification (Accuracy)

$$\text{Accuracy} = \frac{1}{n} \sum_{i=1}^{n} I(\hat{y}_i = y_i) \tag{15}$$

VI RESULTS AND ANALYSIS

For analysis purposes, we will use the research questions from Section II. Our inquiry focuses on the dataset (see Section V).

- RQ-1 Which insights can be derived from the dataset using Unsupervised clustering models?

  Unsupervised clustering models provide valuable insights into the dataset. From the heat map analysis, it is evident that as countries become wealthier, the average income of their citizens also increases. Moreover, countries with higher healthcare expenditures tend to have longer life expectancies for their citizens. This suggests a positive correlation between (income) and (gdpp), as well as between (lifeexpec) and (health) (Fig. 1).

  Additionally, the clustering analysis reveals strong negative correlations. For instance, as countries become richer, their infant mortality rates (childmort) tend to decrease. Similarly, countries with higher inflation rates (inflation) tend to have lower life expectancies (lifeexpec) for their citizens. These findings highlight the complex interplay between economic indicators and health outcomes, indicating the importance of addressing both aspects in policymaking and development strategies (Fig. 1).

  Hierarchical Cluster Analysis (HCA) further elucidates the underlying structure within the dataset by grouping objects based on their similarity or dissimilarity. In this context, HCA identifies two distinct groups, with countries like Qatar, Luxembourg, Norway, and Switzerland standing out in one of them. The resulting dendrogram, a tree-like diagram, provides a hierarchical organization of the clusters, offering a visual representation of the relationships among countries based on the selected indicators (Fig. 2).

- RQ-2 Which Supervised classification models provide the best performance for predicting outcomes in the dataset?

  To determine the best performing classification models for predicting outcomes in the dataset, we evaluated the dataset with one metric, the accuracy. The results indicate that different models excel in different aspects of prediction performance.

  For global accuracy, the Random Forest Classifier achieved the highest score of 0.9804, indicating that it correctly predicted the majority of the outcomes in the dataset. This suggests that the Random Forest Classifier is highly effective in accurately predicting both positive and negative outcomes (Fig. 3).

  In terms of classification accuracy, the AdaBoost Classifier had the lowest score of 0.9216, indicating that it made the most incorrect predictions overall. This suggests that AdaBoost has a relatively lower performance in minimizing prediction errors in this precise case (Fig. 3).

  For accuracy, KNeighborsClassifier, Decision Tree Classifier, and XGBClassifier achieved identical scores of 0.9608, indicating that they were able to predict the outcomes with a high level of accuracy. This suggests that these classifiers are particularly effective in identifying outcomes in the dataset (Fig. 3).

  H0similar: The performance evaluations of the four classification models are similar. *H1similar*: The performance evaluations of the four classification models are not similar. For the four classification models, the p-value $= 0.0625 > \alpha = 0.05$. This indicates that there is no statistically significant difference in the performance evaluations of these models. Thus, our approach demonstrates consistency across the models.

- RQ-3 Which Supervised regression classification models provide the best performance for predicting outcomes in the dataset?

To determine which regression models, provide the best performance for predicting outcomes in the dataset, we evaluated several metrics including R-squared values for each model. The results indicate that different models excel in different aspects of prediction performance.

For R-squared, which measures the proportion of the variance in the dependent variable that is predictable from the independent variables, the KNeighbors Regressor performed the best with a score of 0.9352. This indicates that the KNeighborsRegressor model has the highest explanatory power in predicting outcomes (Fig. 4).

In terms of performance, the Ridge Regressor also demonstrated strong results with an R-squared score of 0.9220, making it another highly effective model in explaining variance in the dataset (Fig. 4).

The Random Forest Regressor and XGB Regressor models also performed well with R-squared scores of 0.9149 and 0.9117, respectively, indicating that they are capable of explaining a substantial amount of the variance in the dataset (Fig. 4).

Overall, the KNeighborsRegressor emerged as the top performer in terms of R-squared, while other models like the Ridge Regressor, Random Forest Regressor, and XGB Regressor also exhibited strong performance. These findings suggest that these models are particularly suitable for predicting outcomes in the dataset, each offering unique strengths in terms of prediction accuracy and variance explanation (Fig. 4).

H0similar: The performance evaluations of the six regression models are similar. *H1similar*: The performance evaluations of the six regression models are not similar. For the six regression models, the p-value $= 0.0156 < \alpha = 0.05$. This indicates a statistically significant difference in the performance evaluations of these models. Therefore, our approach shows variability across the models.

- RQ-4 Which countries are in the direst need of aid?

  To determine which countries are in the direst need of aid, we performed clustering using KMeans on the dataset. The clustering analysis revealed a group of countries that exhibited socio-economic indicators indicating significant need for assistance (Fig. 5).

  These countries are identified as being in the Cluster High Need for Assistance based on the clustering analysis, which highlighted their socio-economic challenges. Efforts to aid and support these countries could help alleviate poverty and improve overall well-being for their populations (Fig. 5).

VII CONCLUSIONS

In conclusion, our analysis of the dataset using unsupervised, supervised classification and regression models has provided valuable insights into predicting outcomes and identifying countries in high need of assistance. For classification models, Random Forest Classifier emerged as the top performer with an accuracy of 0.9804, indicating its effectiveness in accurately predicting outcomes. In regression models, the KNeighborsRegressor model exhibited the best performance with the highest R-squared performance with 0.9352, suggesting its ability to provide accurate predictions and explain variance in the dataset.

Our findings have identified 48 countries in high need of assistance, as determined by clustering analysis using KMeans. These countries exhibit socio-economic indicators that indicate significant challenges and necessitate urgent assistance. This project is crucial as it provides a data-driven approach to prioritize aid and resources to countries most in need, ensuring that efforts are targeted effectively to alleviate poverty and improve the well-being of vulnerable populations.

Overall, this article contributes to the field by showcasing the efficacy of classification and regression models in predicting outcomes and identifying countries in need. By highlighting the best-performing models and identifying countries in high need of assistance, this research provides a valuable tool for policymakers and humanitarian organizations to allocate resources effectively and make informed decisions to address global challenges.

VIII ACKNOWLEDGMENTS

This research was supported by LEAD University (No. 2024-ll BCD7213)

IX REFERENCES

[1] V. Grossi, F. Giannotti, D. Pedreschi, P. Manghi, P. Pagano, and M. Assante, “Data science: a game changer for science and innovation,” Int. J. Data Sci. Anal., vol. 11, no. 4, pp. 263–278, 2021.

[2] Z. Sun, Z. Wu, and K. D. Strang, “Big data-driven socioeconomic development: An interdisciplinary approach,” in Handbook of Research on Driving Socioeconomic Development With Big Data, Hershey, PA: IGI Global, 2023, pp. 1–21.

[3] S. Singh, “Here’s how impact data science can help solve global crises,” World Economic Forum, 20-May-2022. [Online]. Available: https://www.weforum.org/agenda/2022/05/impact-data-science-can-help-solve-global-crises-but-only-if-we-deploy-it-on-the-front-lines/. [Accessed: 06-Apr-2024].

[4] Sabouri, Z., Gherabi, N., Amnai, M. “Comparative Study of Supervised Regression Algorithms in Machine Learning”, vol 826. Springer, Cham. 2024.

[5] Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., Aljaaf, A.J. “A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science.” Unsupervised and Semi-Supervised Learning. Springer, Cham. 2020.

[6] Sohan, Md. Evon Shahriar, et al. ”Optimizing development aid allocation: A data-driven approach using unsupervised machine learning and multidimensional indices.” World Journal of Advanced Research and Reviews 19.3 (2023): 1393-1409.

[7] Baboshkin, P., Yegina, N., Zemskova, E., Stepanova, D., et Yuksel, S. (2021). Non-classical approach to identifying groups of countries based on open innovation indicators. Journal of Open Innovation: Technology, Market, and Complexity, 7(1), 77.

[8] J. Q. Jerin, N. K. Khan, S. Biswas, and N. Sharmin, ”Comparative Study of Clustering Algorithms: Scenario Based on Boston Crime Dataset,” in 2023 26th International Conference on Computer and Information Technology (ICCIT), Cox’s Bazar, Bangladesh, 2023, pp. 1-6.

[9] A. Jan van der Veen, ”Analytical method for blind binary signal separation,” Proceedings of 13th International Conference on Digital Signal Processing, Santorini, Greece, 1997, pp. 399-402 vol.1.

[10] T. Chauhan, S. Rawat, S. Malik and P. Singh, ”Supervised and Unsupervised Machine Learning based Review on Diabetes Care,” 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2021, pp. 581-585.

[11] A. Morton, E. Marzban, G. Giannoulis, A. Patel, R. Aparasu and I. A. Kakadiaris, ”A Comparison of Supervised Machine Learning Techniques for Predicting Short-Term In-Hospital Length of Stay among Diabetic Patients,” 2014 13th International Conference on Machine Learning and Applications, Detroit, MI, USA, 2014, pp. 428-431.

[12] Y. He, et al., ”Outlier Detection Methods for High-Dimensional Data,” in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2659-2668, 2018.

[13] J. Olmedo, et al., ”Data Cleaning and Preprocessing for Machine Learning,” arXiv preprint arXiv:2004.08982, 2020.

[14] G. Chandrasekaran, S. Lakshminarayanan, S. Sankaranarayanan, S. Mehta, and

K. Chaudhuri, ”k-Nearest Neighbors: A Comprehensive Guide,” in Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 377-385, PMLR, 2018.

[15] A. K. Jain, ”Data clustering: 50 years beyond K-means,” Pattern Recognition, vol. 43, no. 3, pp. 651-666, 2010.

[16] C. Strobl, et al., ”Bias in Random Forest Variable Importance Measures for Classification and Regression,” The Journal of Machine Learning Research, vol. 8, pp. 1117-1133, 2007.

[17] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning with Applications in R. Springer Science et Business Media, 2013.

[18] J. Murillo-Morera, C. Quesada-López, C. Castro-Herrera, and M. Jenkins, ”An Empirical Evaluation of NASA-MDP Data sets using a Genetic Defect-Proneness Prediction Framework,” in 2016 IEEE 27th International Symposium on Software Reliability Engineering (ISSRE), Ottawa, ON, Canada, 2016, pp. 1-10.

X APPENDIX

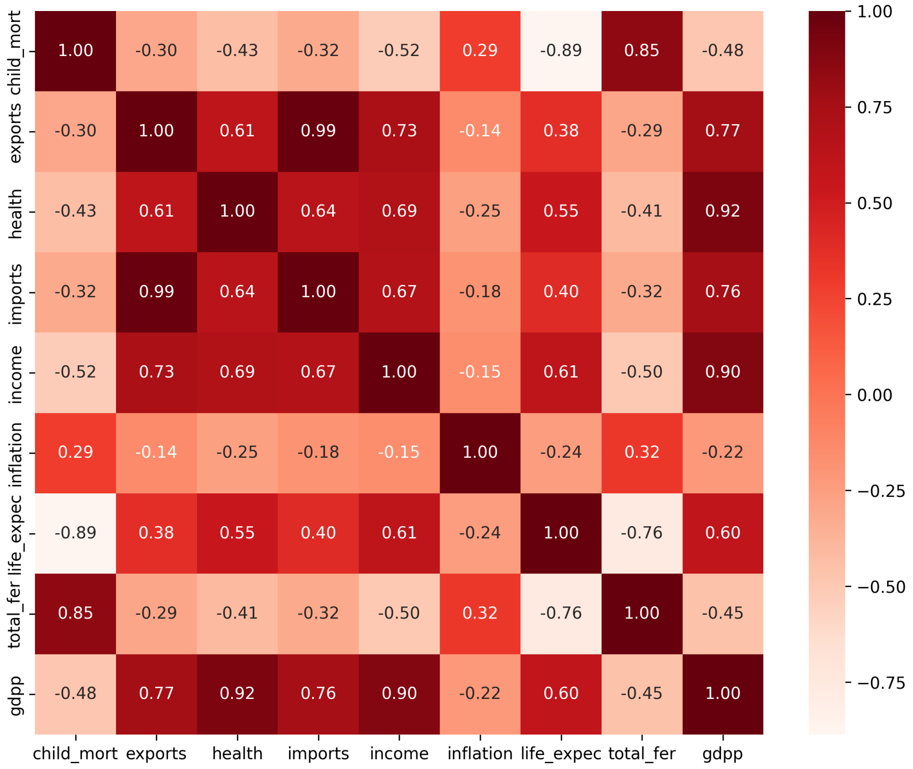


Figure 1: Heat map of Country-data.

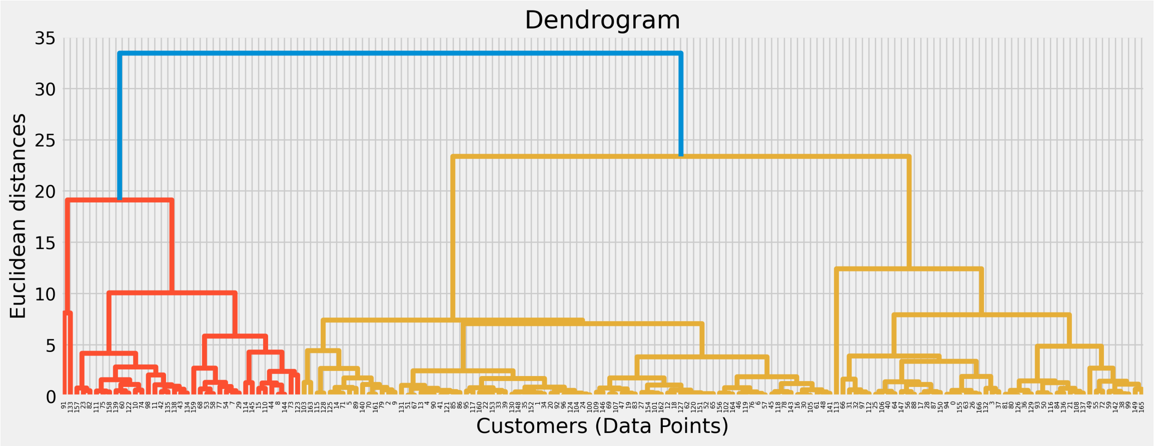
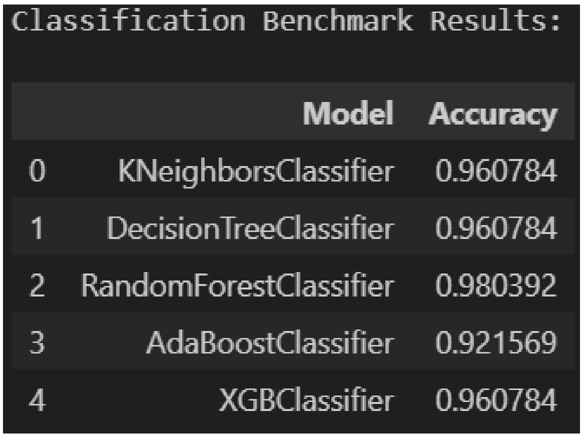


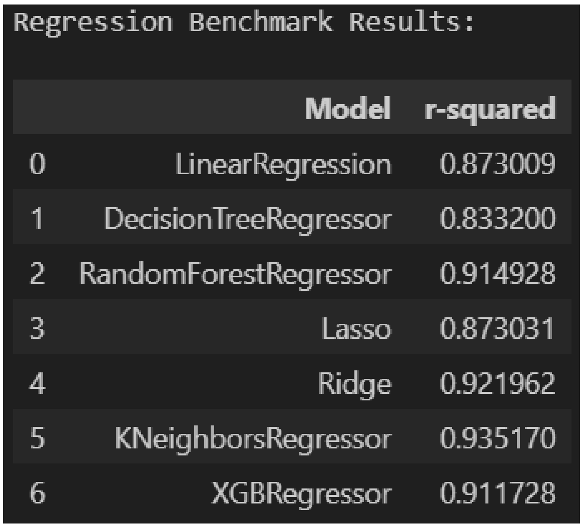
Figure 2: HCA Dendrogram of Country-data.



Classification Benchmark Results:

	Model	Accuracy
0	KNeighborsClassifier	0.960784
1	DecisionTreeClassifier	0.960784
2	RandomForestClassifier	0.980392
3	AdaBoostClassifier	0.921569
4	XGBClassifier	0.960784

Figure 3: Classification Benchmark of Country-data.



Regression Benchmark Results:

	Model	r-squared
0	LinearRegression	0.873009
1	DecisionTreeRegressor	0.833200
2	RandomForestRegressor	0.914928
3	Lasso	0.873031
4	Ridge	0.921962
5	KNeighborsRegressor	0.935170
6	XGBRegressor	0.911728

Figure 4: Regression Benchmark of Country-data.

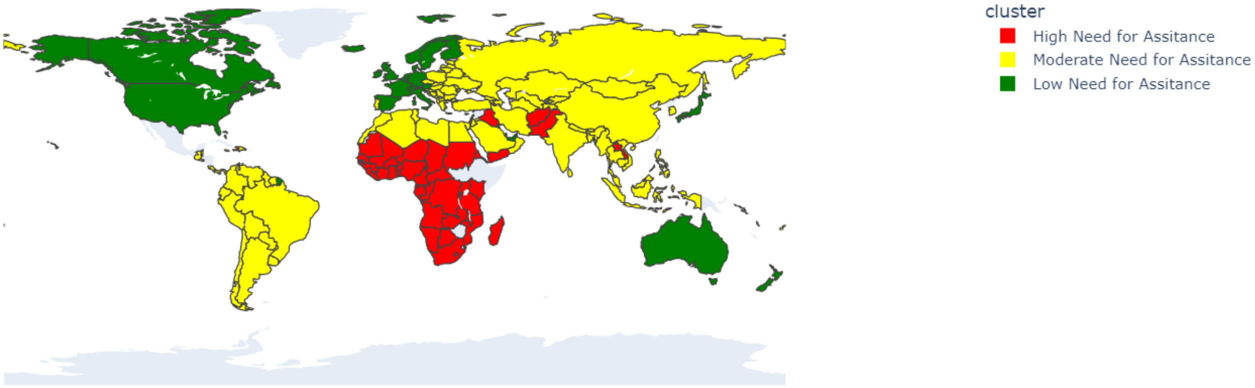


Figure 5: Clusters of Country-data countries.

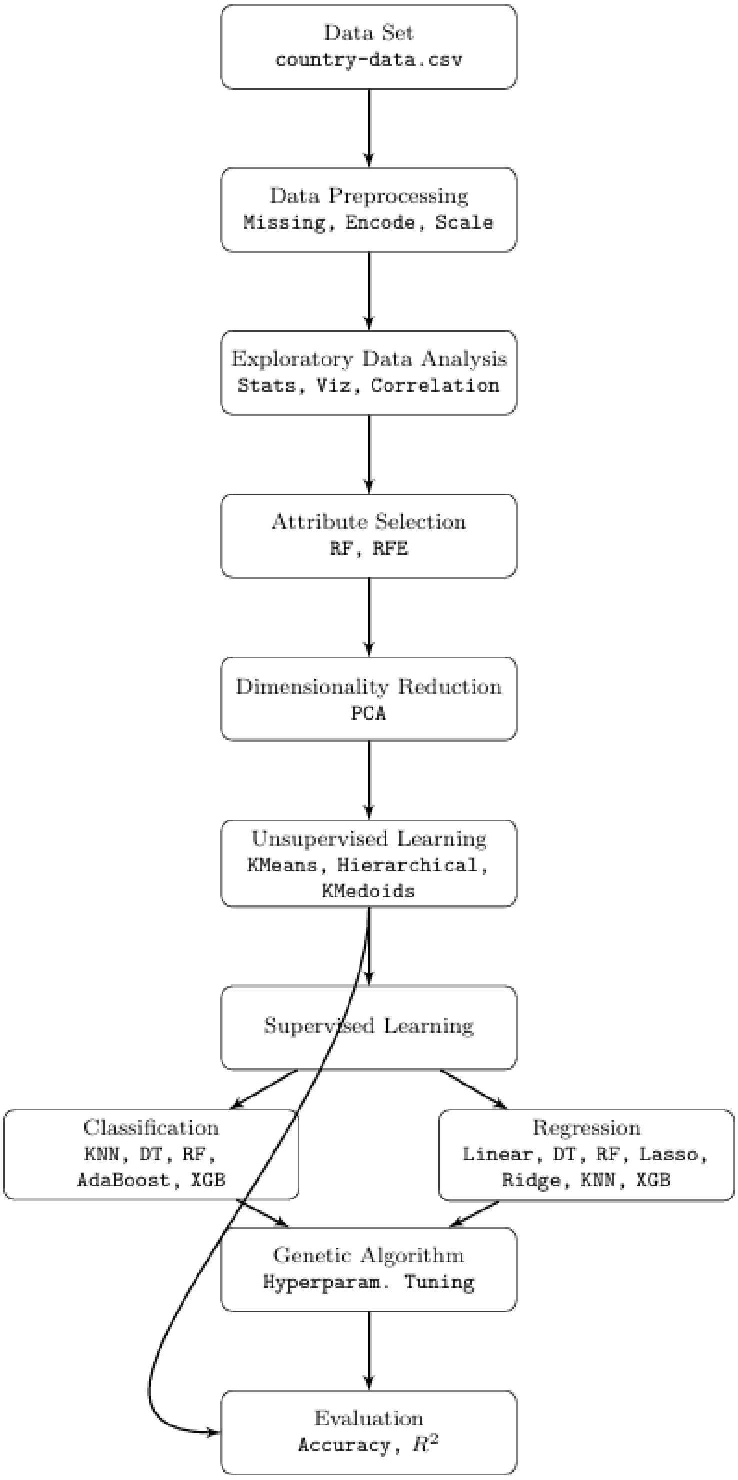


Figure 6: Framework diagram.